Big Data Analytics - Hadoop 2.0
4-Day Practical Training

Why Big Data and Hadoop 2.0

The use of Big Data is becoming a crucial way for leading companies to outperform their peers. It enables us to tap into large information flow where data about products and services, buyers and suppliers, consumer preferences and intents can be captured and analyzed. Therefore nurturing Big Data Analytics expertise plays an essential role in developing new products or services.

Who should attend this Hadoop 2.0 training

This training is ideal for professionals who wish to implement Big Data Analytics by using Hadoop framework.

What you will learn on Hadoop 2.0

The course equips participants to work on the Hadoop environment with ease and learn vital components such as Flume, Hive, Pig, Map Reduce and other advanced concepts like Hadoop 2.0: HBase, ZooKeeper and Sqoop.

The course consists of:

4 Days Classroom training
28 Hours of Real Time Industry based Projects
15 Hours of Lab Exercises with proprietary VM
3 Hours of Doubt Clarification Session
2 Projects with Unique Data Sets Included
Industry Specific Projects on the Retail and Telecom Sectors
Java Essentials for Hadoop Included
Hadoop Installation Procedure Included
Hadoop Deployment and Maintenance Tips
Packed with Latest & Advanced modules like Yarn, Flume, Oozie, Mahout, and Chukwa
Tips and Techniques to clear Certification Exam by the Trainer

About Go Training

Go Training applies effective pedagogical methodologies that demonstrate case studies and hands-on practical skills, in addition to explaining clearly how things work in principle. Every course that we conduct is delivered by a subject matter expert who holds the academic qualification and working experience in that specialization. On the days when they are not teaching, our trainers work on consultancy projects and technical deliveries. Their work has received numerous recognition and awards in the industry. Our team of trainers has been invited as keynote speakers at numerous international conferences, and as principal consultants for various industries.

About the Instructor

Tirumala Rao Dumpala is the Senior IT Consultant who has 7 years of versatile experience in designing, developing and supporting IT applications behind him, which includes three and a half years in Big Data, Text Mining, Analytics, java, Data Mining and Web Development. He graduated with a degree of Master of computer Applications at JNTU. He is currently affiliated with Cap Gemini (HP Client), dealing with HP Real time Analytics and Data processing project with live data volumes around 0.5 TB to 2 TB. His previous working experience includes also position with held with HCL, dealing with the Aadhaar project which developed the unique identification number generation and authentication system. He was also the technical lead of TCS, he was involved in the development of the Business Intelligence and Reporting system which can handle 2 TB to 22 TB of data.

Date: 15-18 March 2016
(Tuesday - Friday)
Time: 0900 - 1700
Venue: Suite 2B-21-1, Level 21, Block 2B, Plaza Sentral, Jalan Stesen Sentral 5, KL Sentral, 50470 Kuala Lumpur, Malaysia.

HRDF Claimable

Big Data Analytics - Hadoop 2.0
4-Day Practical Training

Course Outline

Day 1

Introduction to Big Data
- What is Big Data, Examples of Big Data, Use cases of Big Data
- What is Hadoop, History of Hadoop, Where Hadoop is being used
- Problems with Traditional Large-Scale Systems and Need for Hadoop
- Understanding distributed systems and Hadoop

Software Installation
- Pre-requisites, Understanding Hadoop Configuration Files
- Setup Single Node Hadoop Cluster, The Command-Line Interface

Hadoop Ecosystem
- HDFS, MapReduce, Hive-Introduction, Sqoop-Introduction
- Pig-Introduction, HBase-Introduction, Flume-Introduction
- Spark-Introduction, Oozie-Introduction

Understanding Hadoop Distributed File System (HDFS)
- Understanding Hadoop / HDFS Architecture
- Hadoop Components - HDFS, Map reduce
- Name Nodes and Data Nodes, Hadoop 2.0 Architecture
- Running Hadoop, Web-based cluster UI-Master UI, Map Reduce UI

Hands-On Exercise: HDFS Commands
- Basic HDFS commands

Understanding Map Reduce
- How Map Reduce works, Data flow inMapReduce
- Map operation, Reduce operation, MapReduce Driver Class
- Running your First Program, Split, Record Reader(RR), Sorter
- Shuffler and Partitioner, Combiner in-depth, Distributed Cache
- Writing first MapReduce Drivers, Mappers and Reducers in Java with eclipse, Code Walkthrough, Error Handling in MapReduce
- Map Reduce Job Execution Flow In-Depth

Day 2

Hands-On Exercise: Map Reduce
- First Map Reduce Program with Basic Word Count
- Calculate Aggregation for Structured Data
- Handling Unstructured Data, Processing Fixed Length Values

Hive
- Introduction to Apache Hive, Hive architecture, Installing Hive
- Getting Data into Hive, Hive-HQL & Query Execution
- Working with WHERE Clause, Partitions in Hive (Static and Dynamic)
- Performing JOIN Operation in Hive (Map and Reducer Side Joins)
- Compression in Hive (ORC), Executing hive queries in real time

Hive
- Hive Query Hands On Exercise
- Loading Data, Sample Query with WHERE and JOIN, Partitions

Sqoop
- Installing & Configure Sqoop, Import RDBMS data to Hive using Sqoop
- Export from Hive to RDBMS using Sqoop, Incremental Load

Hands-On Exercise:Sqoop
- Import Data from RDBMS to HDFS and Hive
- Export Data from HDFS or Hive to RDBMS

Day 3

Pig
- Introduction to Apache Pig. Install Pig, Pig Architecture
- Pig Latin – reading and writing data using Pig, Parameter Passing with Pig, UDFS in PIG, Managing Multiple Pig Scripts in Real-Time Case
- Executing Pig Scripts in Real-Time Projects

HBASE
- What is HBase, Install HBase, HBase Architecture
- Command line interface Exercise, MapReduce Programs in HBase
- Filters in HBase, HBase – Hive Integration

Hands-On Exercise:HBase
- Hbase command line interface, lading data into Hbase with MapReduce.

Day 4

Spark
- What is Spark & why, Install Spark, Spark Cluster Standalone Mode and UI, Using the Spark Shell, Spark Components
- Spark Streaming Overview, Functional Programming with Spark

RDD
- Resilient Distributed Datasets (RDDs), Key-Value Pair RDDs
- Spark Interface with Scala and Java

HDFS
- RDD Partitions and HDFS Data Locality, MapReduce and Pair RDD Operations, Programming in Spark,
- Example: Streaming Word Count, Creating the SparkContext
- Configuring spark properties, caching overview, distributed persistence
- Other streaming operations, common spark algorithms, iterative algorithms, Building and Running a Spark Application, Logging

Hands-On Exercise:Spark I
- Hands on Examples on Spark Shell, Hands on Spark MapReduce
- Building Spark Application

Flume
- Introduce Flume, Flume Installation
- Flume Components (Agent, Sourse, Channel, Sink, Receiver)
- Flume Configuration with Source to Write Data into File (Local and HDFS), Multiple sources with Flume
- Running Flume Agent, Running Receivers and Test with Sample Data

Hands-On Exercise:Spark II
- Configuring Flume with HDFS sync, Streaming Data to Spark
- Validating Chat Application

Training Course Project
- Real time project by taking real time data, Take the data from different source system like text files, CSV files, RDBMS, Loading the data into Hadopp & develop analytics solutions using MapReduce, HIVE& PIG

We will conduct an evaluation test at the end of the course if time is permissible.

Go Training
wholly owned by iRadar Sdn Bhd
HRDF Approved Training Provider (Category A)

No. 36, Jalan IMJ 1, Taman Industri Malim Jaya, 75250 Melaka, Malaysia.
t +606 336 6016
f +606 252 3059
w www.gotraining.com.my
[f] fb.com/gotraining.com.my
[in] linkedin.com/company/gotraining

To register, please contact:
m +6010 663 1852
e yiwei@gotraining.com.my